Text Based Web Image Retrieval System by Using Latent Semantic Indexing (LSI)

Swe Swe Lwin, Khin Mar Myo
University of Computer Studies, Mawlamyine, Myanmar
sweswelwin.tym@gmail.com

Abstract

This paper describes the use of latent semantic indexing (LSI) in text-based image retrieval system. Due to the rapid development and popularity of World Wide Web (WWW), users have to face a variety and large number of web pages and images that waste time to search and browse. To alleviate this difficult, more efficient image retrieval system is needed to extract the required image from the collection of web pages. Therefore many image retrieval systems have been developed and used on the web. Among them, text-based image retrieval is a kind of image Meta search that based on associated metadata. Our proposed text-based image retrieval can retrieve both images alone and combined of images and associated alphanumeric text.

1. Introduction

Nowadays, various types of image are abandoning on the website. Therefore, image searching occur some difficulties when surfing the web. To overcome this problem, Image mining uses methods from computer vision image processing, image retrieval, data mining, machine learning, database and artificial intelligence.

Many web search engines either commercially such as Google Image Search, Yahoo Image Search, Lycos, Alta Vista photo finder, Ditto, art.com, Amico or as research prototypes such as Image Rover, Web Seek, Atlas WISE, PicToSeek, and Page Rank for Product Image Search have been successfully developed. They also stated that web image search system still

Most of commercial web image search systems, such as Google, Lycos, and AltaVista, only support keyword based search. These systems use surrounding blocks of text to index the corresponding images. Some academic community use weight to represent the semantics of web images. Some system only uses visual features to index and search web images. Example system is the PicToSeek which uses pure content-based web image retrieval system. Other systems make use of some combined models to support web image searches.

First part of this paper, we describe how to develop web image search by applying the keyword-based approach. Then how to conduct the indexing and searching using the Latent Semantic Indexing (LSI) is followed. Because, the LSI method is the one of the dimension reduction methods which can reduce the dimensions of feature vector. Beside these, LSI is a statistical Information Retrieval (IR) method that is capable of retrieving text based on the concepts it contains, not just by matching specific keywords.

The remaining section of this paper is organized as follows. Section 2 presents the motivation of this paper. Section 3 expresses the background theory. Section 4 describes about applying web content management system theory on Blog. Section 5 presents the system implementation of this paper. Section 6 describes about conclusion and future works of this paper.

2. Related Works

The author [15] presented an effective approach to and a prototype system for image retrieval from the Internet using web mining. One of the key ideas in that approach is to extract the text information on the web pages to semantically describe the images. The text description is then combined with other low-level image features in the image similarity assessment.

Another author [8] firstly stated that image search engines based on text keywords can fetch thousands of images for a given query, but the results may contains noise. Secondly, the author [8] presented a technique that allows a user to refine noisy search results and characterize a more precise visual object class. Their technique is based on semi-supervised machine learning in a novel probabilistic graphical model composed of both generative and discriminative elements. They demonstrated their approach on images of musical instruments collected from Google image search.

The author [12] discussed about recently researches and developments concerning with digital image retrieving and image annotation in the transaction paper of IEEE. The author [6] proposed a keyword propagation method for image retrieval based on a recently developed manifold-ranking algorithm.
The author [3], presented an approach for image retrieving and browsing based on Scenique image. Scenique is based on a multi-dimensional model, where each dimension is a tree-structured taxonomy of concepts, also called semantic tags that are used to describe the content of images.

3. Background Theory

The related literature is summarized in this section. The main concerns are image retrieval and Latent Semantic Indexing (LSI).

3.1. Image Retrieval

An image retrieval system is a computer system for browsing, searching, and retrieving images from a large database of digital images. Most traditional and common methods of image retrieval utilize some method of adding metadata such as captioning, keywords, or descriptions to the images so that retrieval can be performed over the annotation words. Manual image annotation is time-consuming, laborious and expensive; to address this, there has been a large amount of research done on automatic image annotation [1].

Image search is a specialized data search used to find images. To search for images, a user may provide query terms such as keyword, image file link, or click on some image, and the system will return images "similar" to the query. The similarity used for search criteria could be Meta tags, color distribution in images, region/shape attributes, etc [7].

The author [5] stated that there are three approaches to search the image.

- *Image meta search* search of images based on associated metadata such as keywords, text, etc.
- *Content-Based Image Retrieval (CBIR)* the application of computer vision to the image retrieval. CBIR aims at avoiding the use of textual descriptions and instead retrieves images based on similarities in their contents to a user-supplied query image or user-specified image features.
- *List of CBIR Engines* list of engines which search for images based image visual content such as color, texture, shape/object, etc.

3.2. Text-Based Image Retrieval

Unlike image retrieval from a fixed database, where each image is treated as an independent object, for image retrieval over the Web each image comes along with a host page, which contains a great deal of relevant information about the image. In general, for most of the images on the Web, their content is more or less related to the content of the host pages. For example, a photo of Mars is found more likely from a page talking about space and planets than from a page talking about pop-music. Therefore, we can use not only the image file names but also the page titles and text terms around the images to index and retrieve the images. This actually makes text-based image retrieval more efficient over the Internet than over a database since manual annotation is no longer required.

The text-based approach is also significantly more efficient than the CBIR system on the Internet, in terms of computational cost as well as image transmission and storage cost. For the CBIR system, it is next to impossible to transmit, store, and compute content features for the unlimited amount of images on the Internet. On the other hand, text document retrieval over the Internet has become a routine task for a commercial web search engine. Retrieving images based on text is even simpler since only the portion of the document around the image needs to be searched. Due to the low cost, text-based approach can retrieve significantly more relevant images over the Internet, therefore gives a much higher recall rate than the CBIR approach.

However, accompanying the relevant search results, there could be a large number of irrelevant search results, i.e. the precision of the text-based search can be low. In many situations, a few words cannot precisely describe the image content, and many words have multiple meanings. For example, the query term *sun* may retrieve photos of the Sun or the logos of SUN Microsystems Company. Here, the definition of relevancy depends on the interest of the user. With a low-precision retrieval result, a user may soon lose patience flipping through dozens of pages of images that contain many irrelevant images.

3.3. Latent Semantic Indexing

LSI is being used in a variety of information retrieval and text processing applications, although its primary application has been for conceptual text retrieval and automated document categorization.

LSI uses a term-document matrix to identify the occurrence of terms within a set of

documents, applies term weighting based on term frequencies to reflect the fact that some terms are more important than others in a body of text, and then performs a Singular Value Decomposition (SVD) on the matrix to determine patterns in the relationships between the terms and concepts used in the documents. LSI uses a mathematical transform technique to reduce the number of dimensions in the term space of the matrix to make it more useable and efficient. One consequence of LSI processing is the establishment of associations between terms that occur in similar contexts. As a result, queries against a set of documents that have undergone LSI will return results that are conceptually similar in meaning to the query even if they don't share a specific word or words with the query.

The detail steps of LSI and how to apply the LSI in our system is detail explained in next Section.

4. Overview of the System Design

The main objective of the system is to develop the web image search system using keyword-based approach. There are three main components in system design: (i) HTML parser, (ii) Text Preprocessor, (iii) Image Matcher.

4.1. HTML Parser

This component is responsible to parse the web page into web page contents heading, title, paragraph and photo, etc based on requirement for next phase. In this system, heading field, title field, and paragraph field are only interested for indexing and searching. The example HTML page interested tag and tag value is illustrated in Figure 1.

```
<html>
    <head>
        <title>Sunset</title>
    </head>
        <body>
        <img src="sea and sunset.jpg">
        <img src="seasunset.jpg">
        <h2> Beautiful Sunset</h2>
        <p>
        In my country has
        beautiful sunset in many places.
        </p>
        </body>
</html>
```

Figure1. Example HTML Page

4.2. Text Preprocessor

The sentence and phrases getting from phase 1 is preprocessed in this phase. Sub processes of this component are word tokenizing, stop-word removing and token stemming.

In tokenizing, sentences and phrases are tokenized and sorted to ease in weight measure. The un-affected words (stop-words) in similarity measure, such as article, pronoun, and adjective are removed in stop word removing process. The normalized forms of words (removing pre-fix/post-fix) are performed by stemming process.

After that the tokens getting from respective document and images are recorded and indexed in respective table for next process.

4.3. Image Matcher

This component is responsible to search the image based on index table and LSI. There are two main steps in this process. The first one is to find the web document using LSI and the second one is to retrieve the images from the corresponding web document from indexing table.

The following is the example calculation of LSI.

Step 1: Score term weights and construct the term-document matrix A and query matrix:

d1: Beach homes
d2: Beach houses
d3: Beach bungalow

Step 2: Decompose matrix A matrix and find the U, S and V matrices, where

$$\mathbf{A = USV^T}$$

$$U = \begin{bmatrix} 0.8660 & -0.0000 & -0.0000 & -0.5000 \\ 0.2887 & -0.4082 & -0.7071 & 0.5000 \\ 0.2887 & 0.8165 & 0.0000 & 0.5000 \\ 0.2887 & -0.4082 & 0.7071 & 0.5000 \end{bmatrix}$$

$$V = \begin{bmatrix} 2.0000 & 0 & 0 \\ 0 & 1.0000 & 0 \\ 0 & 0 & 1.0000 \\ 0 & 0 & 0 \end{bmatrix}$$

$$V = \begin{bmatrix} 0.5774 & 0.8165 & 0 \\ 0.5774 & -0.4082 & 0.7071 \\ 0.5774 & -0.4082 & -0.7071 \end{bmatrix}$$

$$V^T = \begin{bmatrix} 0.5774 & 0.5774 & 0.5774 \\ 0.8165 & -0.4082 & -0.4082 \\ 0 & 0.7071 & -0.7071 \end{bmatrix}$$

Step 3: Implement a Rank 2 Approximation by keeping the first columns of U and V and the first columns and rows of S**.**

Step 4: Find the new document vector coordinates in this reduced 2-dimensional space.

d1 (0.5774 0.8165)
d2 (0.5774 -0.4082)
d3 (0.5774 -0.4082)

Step 5: Find the new query vector coordinates in the reduced 2-dimensional space.

$q = q^T U_k S_k^{-1}$ k=2

Step 6: Rank documents in decreasing order of query-document cosine similarities**.**

The following is the example calculation for cosine similarity measure for the above example. Let the query term is "beach",

$sim(q,d) = \frac{q*d}{|q||d|}$

sim (q,d1)= 0.9429
sim (q,d2)= 0.3336
sim (q,d3)= 0.3336

According to similarity value, the document d1 and d2 their corresponding images are retrieve from second portion of this process. Finally these images are provided to user as a search results.

The system design and their components are described in Figure 2.

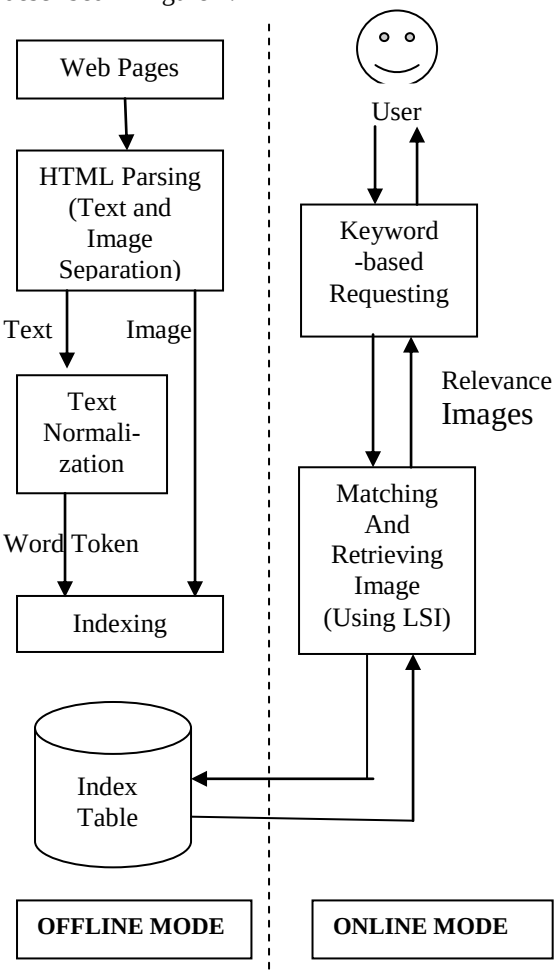


Figure2. Overview System Design

5. Implementation and Results

A prototype system is developed based on C#.Net and SQL Database platform to evaluate the effectiveness and accuracy of the proposed system design.

5.1. Experiment

For offline mode, appropriate amount of web pages concerning with weather forecasting and world climate change are collected. Then this HTML page becomes the sources of our system. The required processes are pre-conducted as describe in system design in offline mode.

For online mode, two types of users are involved in this system*: normal user* who search the images and *admin* who have facility both to search image at the online mode and to update the site at the offline mode. The main interface for this system is illustrated in Appendix 1.

For image search, user only needs to enter the word or phrase or sentence for image. Let the user query term is "beach", the result page is obtained as described in Appendix 2.

As another facility, users have to choose one of two forms of search results. The first form is image alone, the second form is page links. The result for second form can be seen in Appendix 3. Search result of second type is very similar with Google. It contains both text and images content.

5.2. Experimental Results

The accuracy for this system is measured based on the precision.

- Precision

  Precision (P) : $P = \frac{TP}{TP + FP}$

- Recall

  Recall (R) : $R = \frac{TP}{TP + FN}$

In this experiment, the precision is measured with the different threshold value on two forms of search. The rates of containing un-relevance images results are more in image alone (Appendix2) form than link (Appendix3) form. When we conduct one query with two forms of search, the results in the form of URL link is more accurate them image result as shown in Figure 3 and Figure 4.

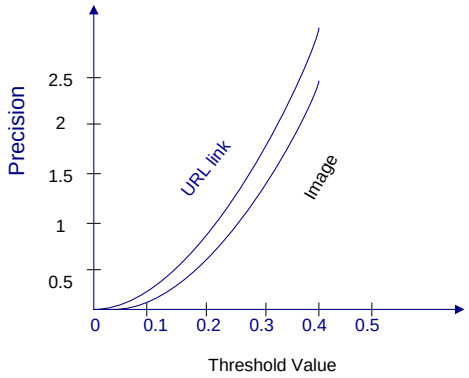


Figure 3. The results of Precision

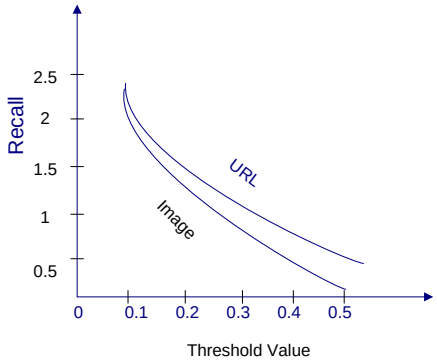


Figure 4. The results of Recall

6. Conclusion

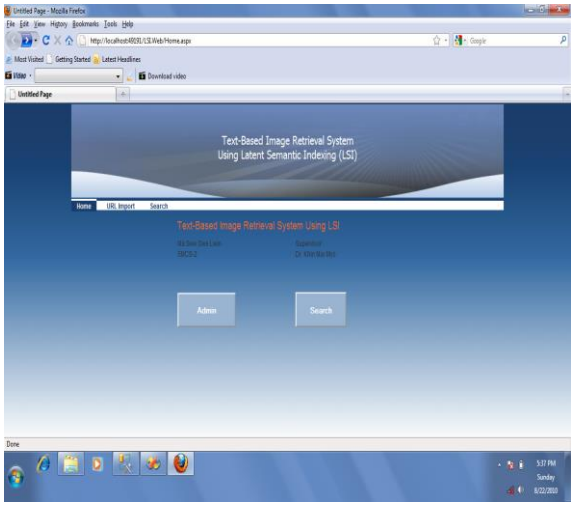
The main purpose of this paper is to provide an efficient and truly realizable approach for WWW based image retrieval. This system performs a text-based meta-search to obtain an initial image set with relatively high recall rate and low precision rate. Then the image content based processing is employed to produce a much more relevant output. There are three ways to combine the text-based method and the visual content-based method: use the text-based method first; use the visual content-based method first; use the two methods at the same time. The key to the success of our system is using Latent Semantic Indexing (LSI). By using LSI and low-cost text-based method to collect as many relevant images as possible over the Internet at a very low cost and to reduce dimension of document matrix. Then the high-precision and high-cost visual based method is used to improve the relevance precision on a significantly smaller image set. Experimental results show that, even with the simplest image feature and clustering algorithm the system achieves promising results. More extensive experiments are needed to test the system. We expect improved performance using more elaborate visual features to describe the image content.

REFERENCES

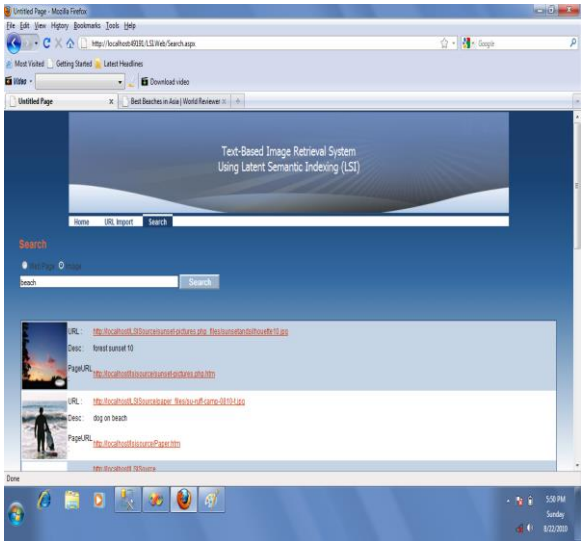
[1] S. E. Arnold, "Right-sizing Content Management", Arnold Information Technology.
[2] Art Technology Group, "ATG White Paper: Understanding Web Content Management Systems", February 13, 2005.
[3] Ilaria Bartolini, "Multi-dimensional Keyword-based Image Annotation and Search "DEIS University of Bologna, Italy.
[4] G. E. Bock, Sr. VP and Sr. Consultant, Patricia Seybold Group, "Content Management Framework", Boston.
[5] M. J. Halm and M. Pelikan, "Enterprise Content Management Systems: Beyond Digital Asset Management and Web Content Management Systems", May 2002, Penn State.
[6] Hang hang Tong1, †, Jingrui H,"Manifold-Ranking Based Keyword Propagation for Image Retrieval"
[7] Y. Kang, Y. Kidawara, Y. Kwon and K. Tanaka, "Implementation of the Web-based Local Blog System on Digital Map", Proceedings of International Conference on Internet Information Retrieval 2005, Koyang City, Korea.
[8] Nick Morsillo, Chris Pal, Randal Nelson, "Mining the Web for Visual Concepts"
Univ. of Rochester Comp. Sci. Dept. Technical ReportTR-2008-931March18,2008
[9] P. Kennedy, "Manifestations of metadata: from Alexandria to WCMS, the old is new again", January 2007.
[10] L. Merker, "How to Evaluate Web Content Management Solutions for Higher Education… and Avoid Overspending", Omni Update, Inc.
[11] J. Subrahmanyam, "Future Trends of Content Management Systems (CMS) for e-Learning: A Tool Based Database Oriented Approach", Hyderabad, India.
[12] James Z. Wang, "Real-World Image Annotation and Retrieval: An Introduction to the Special Section", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 30, NO. 11, NOVEMBER 2008
[13] A. I. Wasserman, "Principles for the Design of Web Applications", Center for Open Source Investigation (COSI), Carnegie Mellon West, USA.
[14] X. Zeng and S. T. Harris, "Blogging in an Online Health Information Technology Class", Blogging in an Online Health Information Technology Class..

[15] Zheng Chen, Liu Wenyin, Feng Zhang, Mingjing Li, Hongjiang Zhang," Web Mining for Web Image Retrieval"

Appendix1.



Appendix2.



Appendix3.